DOCUMENT RESUME

ED 059 748 LI 003 487

TITLE	Guidelines for Establishment and Development of Multilingual Scientific and Technical Thesauri for Information Retrieval.
INSTITUTION	United Nations Educational, Scientific, and Cultural Organization, Paris (France).
REPORT NO	SC-WS-501
PUB DATE	30 Dec 71
NOTE	23p.; (0 References)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Development; Guidelines; *Information Retrieval; Lexicography; *Multilingualism; *Thesauri
IDENTIFIERS	*Scientific and Technical Information

ABSTRACT

The present guidelines, preliminary in character, are an attempt to facilitate the development and application of a standardized method of multilingual thesaurus construction and to reduce the number of arbitrary variations in thesaurus techniques. In the preparation of these guidelines the approach has been theoretical rather than practical. Following a definition of the multilingual character of a thesaurus, the guidelines list a set of initial conditions stating various contexts under which thesauri can be established. Then the phases through which the task can be completed are enumerated in chronological order: Establishment of first draft of target thesaurus, Test and checks, Establishment of a stabilized target thesaurus, and Updating and maintenance. (Author/SJ)

SC/WS/501
PARIS, 30 December 1971
Original: English

UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

GUIDELINES FOR ESTABLISHMENT AND DEVELOPMENT OF MULTILINGUAL SCIENTIFIC AND TECHNICAL THESAURI FOR INFORMATION RETRIEVAL

Photographic reproduction
of the original manuscript

FOREWORD

The present guidelines, preliminary in character, are an attempt to facilitate the development and application of a standardized method of multilingual thesaurus construction and to reduce the number of arbitrary variations in thesaurus techniques.

In the preparation of these guidelines the approach has been theoretical rather than practical; therefore critical reviews of practitioners in particular will be welcome.

Comments and suggestions on these guidelines may be sent to:- Division of Scientific Documentation and Information, Unesco, Place de Fontenoy, Paris 7e.

Table of Content

1. Purpose and Definition

"Recent developments in the methodology of information storage and retrieval and the establishment of new information centers have given rise to the creation of many divergent and incongruent subject indexing vocabularies". This trend has a variety of causes which are presumably to endure.

Some are linked to the evolution of scientific and technical research, showing a well-known and ever-increasing specialization. Others relate to socio-political factors whereby nations devote a greater share of their resources to scientific research and development; there ensues a greater interest, or concern, by public agencies to provide adequate exchanges of information while protecting national and/or "linguistic" interests; and on the other hand one witnesses the de facto establishment of international networks of scientific and technical information by more or less regulated private or semi-private agencies - be they professional, or entrepreneurial organizations.

The resulting creation of information systems, including their own brands of indexing vocabularies is a necessary and welcome development; but it needs to be checked and regulated if the task does not end in defeating its own purpose. For the anarchical proliferation of "networks" sometimes covering the same ground, often restricted by national or linguistic barriers, precludes the establishment of efficient scientific exchanges.

Already, in order to avoid such a situation, a project has been launched by Unesco and ICSU to assist in the creation of a world network of scientific and technical information (UNISIST).

All these reasons militate for an "attempt to lay the basis for

/...

compatibility both at present and in the future, of thesauri that are being elaborated simultaneously in most of the disciplines of science, basic as well as applied".

1.1. Definition of a Monolingual Thesaurus

Unesco has already laid out "Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval". The present text aims at complementing them so as to provide information practitioners with a set of compatible, if not uniform, rules.

This conforms, moreover, with the general philosophy of UNISIST's recommendations "so as to avoid both unwelcome monopolies and unproductive proliferations".

As a follow up of the "Guidelines for the Establishment and Development of Monolingual Thesauri", already cited, the present text will retain two basic traits:

1.1.1. It will rest on the same definition as to what constitutes a thesaurus:

In terms of function, a thesaurus is a terminological control device used to "translate" from the natural language into a system language (information language) as well as to translate the system language back into natural language.

In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which comprehensively covers a specific domain of knowledge.

We shall retain as well the definition of a descriptor as "an authorised and formalized term (word or symbol) in a thesaurus, used to represent unambiguously the concepts of documents and queries."

/...

1.1.2. The Guidelines in both cases will remain general in the sense that,

a) they will not be restricted to any particular field of application, or any particular language - even if some distinctions are made for explanatory reasons,
b) they intend to provide a record of problems encountered, indications as to the chronogical, or logical order in which to tackle them and clues to ways of solving them.

However, the largely empirical character of thesaurus building and translation procedures precludes giving hard and fast solutions, and exact answers to all possible cases. The present Guidelines constitute the first element of a multi-phased effort: phase one will attempt to define a multilingual as opposed to a monolingual thesaurus; it will also advise on the first steps to establish it; later on, when sufficient experience has been gained through practical work, the postulated definition may be amended. The present Guidelines are therefore to be considered as a preliminary step aimed at "triggering" implementation.

1.2. Definition of a Multilingual Thesaurus

Following a Definition of the multilingual character of a thesaurus, the Guidelines will list a set of Initial Conditions stating various contexts under which thesauri can be established. Then the phases through which the task can be completed will be enumerated in chronological order: Establishment of first Draft of Target Thesaurus; Tests and Checks; Establishment of a Stabilized Target Thesaurus; Up-dating and Maintenance. Such a chronological division may not be congruent with a more "logical" order: similar analysis or control procedures may be necessary at different times in the recommended schedule, and they will be mentioned as many times as needed.

/...

A multilingual Thesaurus will allow equivalent indexed representations of documents for any given topic expressed in at least two natural languages. It will allow indexers to index documents in the languages they know, and achieve results. In other words considering some properties of thesauri:

1.2.1. A thesaurus requires a number of explanatory statements in natural language – descriptive introductions, definitions, scope notes, etc. The first condition for a thesaurus to be multilingual is that such statements be expressed in more than one natural language.

1.2.2. The second condition is that the two or more versions of the thesaurus make identical indexing possible in the various linguistic contexts they cover. In other words, descriptors have to be defi: d and used to represent identical topics in documents expressed in different languages so that, eventually, they could be uniquely symbolized.

From this definition it follows that the establishment of a multilingual thesaurus requires two types of operations:
- translations, as regards statements necessarily present in thesauri
- "transformations" of descriptors – i.e. their analysis and synthesis – whenever they denote which are uniquely expressed in one natural language.

2. Initial Condition

Two cases may obtain when a multilingual thesaurus is to be established: either there is no available thesaurus in the field or one monolingual thesaurus is available and can be employed.

2.1. Establishment of a multilingual thesaurus when there is no monolingual thesaurus available

The alternative opened here is either the simultaneous creation of two

/...

or more versions, or a sequential procedure by which a monolingual version is later made equivalent to one or more others. As will be shown, the present Guidelines advise building up successively two drafts of the thesaurus, the first being based on description and analysis, the second being the result of actual operating conditions.

2.1.1. The simultaneous establishment of a first draft of a multilingual thesaurus can be envisaged whenever the following set of conditions are met:

- the field of application calls for the creation of new thesauri: the case arises when a discipline has become recognized as autonomous, or when the need is felt for new mission-oriented information exchanges.
- a group of specialists in the field and in information retrieval techniques with adequate linguistic capabilities can be brought to work as a team.

The first condition implies that the field is not mainly expressed in one particular language as well as regards practitioners as regards any established terminology. If found to be so, it is clear that for the sake of quality, the major language serves as the basis to other versions of the thesaurus.

The second is linked to the theoretical and temporary character of the first draft. Inasmuch as the first draft is removed from actual operating conditions - i.e. actually used for indexing and retrieving corpora of documents - a team of specialists in close contact may succeed in devising correct equivalence between various linguistic contexts. In this sense all the versions of the multilingual thesaurus will have to be subjected to the same tests and to be corrected simultaneously.

2.1.2. The establishment of a second draft of a multilingual thesaurus -

i.e. versions usable in two or more linguistic contexts - should better be based on the existence of one monolingual thesaurus.

The justifications for such a procedure are threefold:
- operational: actual operating conditions - with their concurrent corrections, updating and maintenance procedure - require on-the-spot observations and adjustments. As all these may vary greatly according to local conditions it is unrealistic to expect that instantly uniform conditions can prevail (the exception being, of course, an information system using unique and centralized procedure).
- costs: the long term convening of a team of specialists is both costly and unwieldly.
- development: in the course of expanding mono- or multilingual thesauri to other languages, a single linguistic version is to serve as a basis in order to avoid the appearance of uncontrolled terms.

2.2. Whenever there is available a monolingual thesaurus to serve as a starting point, the present Guidelines will apply.

In any case, various initial conditions may be reduced to one of expanding one monolingual thesaurus to cover one or more linguistic environments. The process will thus be designated in the following manner: the establishment of one source - or basic - thesaurus formulated in a natural language (source language), to a target version of the same, formulated in another natural language (target language).

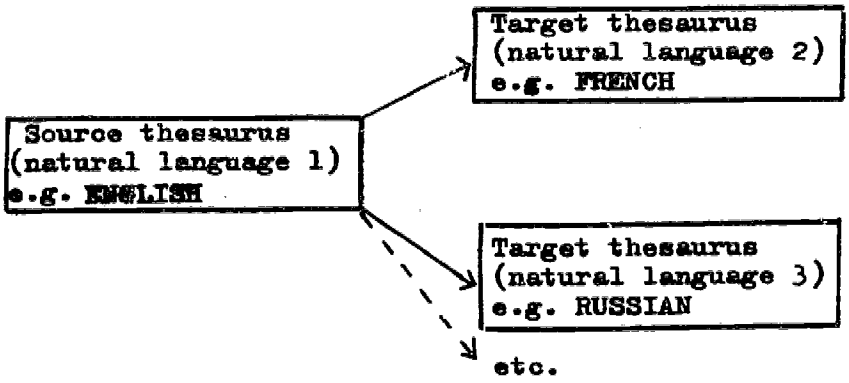

Figure 1

3. Establishment of first draft of a target version

The first draft of the target version will aim at providing a workable image of the source thesaurus, expressed in the target language; the actual use of this draft will make possible the ulterior checks and controls. Three processes will be required here - analysis of source thesaurus, translation, interpretation - which are to be conducted in six different steps.

3.1. Analysis of the source thesaurus

The basic components of the thesaurus are to be analysed and ordered thoroughly so as to enable a progressive re-construction in target language.

A summary of such components, and their presentation in tabular form (see the Descriptive Table of Source Thesaurus) will help determine a logical sequence of operations:

3.1.1. Notation: under this heading comes the actual formulation of the descriptor, whether it be a single natural language word or expression, or any kind of symbolic code (number, alphanumeric sign, etc.). It is not to be confused with the codes eventually used for computer operations.

3.1.2. Semantic Information: collected here are all the informations which help thesaurus users to grasp the meaning of individual descriptors. They fall into two categories: natural language statements such as definitions and scope notes which may be attached to a number of descriptors; and eventually lists of natural language words which the descriptor is supposed to subsume (non-descriptors).

3.1.3. Structural Information: in most cases, inter-relationships between descriptors are exposed as explicit - in the case of hierarchies

/...

- or implicit - whenever "see" and "see also" references are used. There is usually no other information about the true nature of the link established other than formal names - e.g. "broader", "narrower", "related", etc...

Step 1 Provisional descriptor notations are to be established in two forms. The first will make use of target language words or expressions. There is no need to stress at this point the desire or need for homogeneous form of expression - e.g. to select singular vs. plural forms, given parts of speech (substantives, adjectives, verbal forms etc.) - as:

a) the main purpose of such a notation is to give clues as to the descriptor meaning,

b) there are no formal equivalences in this respect between natural languages.

The second notation will be formal (numerical or otherwise) and will serve as a reference to all other forms of the descriptor (preferably for machine operations).

Step 2 All the statements included in the source thesaurus and expressed in natural language will be translated. They will cover as the case may be:

- descriptive introductions dealing with the purpose of the thesaurus, the field covered, the types of documents to be processed, etc.
- all comments about the thesaurus organization, its guiding principles, etc.
- all definitions and/or scope-notes which may be attached to descriptors.

Step 3 Semantic Information is to be interpreted in two ways. Whenever natural language words or expressions are given as meaningful eqivalents to descriptors in the source

/...

a) NOTATION		b) SEMANTIC INFORMATION		c) STRUCTURAL INFORMATION		d) OPERATIONAL INFORMATION	
Term	Symbol	Definitions Scope-notes	Source language equivalences	Type and Nature of related terms (broader, narrower, related, etc.)	Interpretation of relationships	Frequency of use	Date of insertion
Documentation	I.12	The process of storing and retrieving information in all fields of learning	Information Storage and Retrieval	- Dessemination - Information Processing	Part of process ... Set of methods used for ...		

Figure 2: DESCRIPTIVE TABLE OF SOURCE THESAURUS

(Examples drawn from the field of documentation)

thesaurus, target natural language words or expressions should be looked for. At this stage the search can be only empirical - until tests are carried out on actual documents- using all available technical dictionaries, glossaries or terminological tools, of a monolingual nature

Step 4 At this point problems may be raised by the non-overlapping meanings encountered - i.e. by the fact that natural language peculiarities, coupled to specific national practices or schools of thought, generate concepts which have no exact replicas in other contexts. That such a phenomenon should be more prevalent in the "soft" sciences than in the "hard" ones is an empirical observation and is intuitively evident. A more precise description depends upon the setting up of systematic and significant tests.

Two cases may arise: either the source descriptor is more easily defined and represented with the help of two or more target natural language words or expressions; a compound expression - not a single word - is needed to indicate the meaning of the source descriptor; or there comes up, through empirical observations, the necessity for a target descriptor which has no observable equivalent in the source thesaurus: the case arises when a school of thought is uniquely expressed in one natural language, or when national practices covered by the thesaurus have no equivalents. Some descriptors may be present in the source thesaurus and may never be used in the target one. In both cases, target descriptor notations are to be consigned and adequate definitions attached to them. Equivalent descriptors are to be noted down, while a list of potential new descriptors is drawn up. The cases, however, may not be quite as clear cut: there may not obtain a clear one-to-many relationship - e.g. one source descriptor being

equivalent to exactly target ones. Such unclear demarcations will be made apparent in two ways: one by the set of definitions; the second through type of descriptor inter-relationship(s).

Step 5 The interpretation of structural information is to be performed in two main phases. The first is a thorough description and analysis of source descriptors inter-relationships. These may appear in a formal way, as shown by their location in hierarchies, for instance. In which case lists may be drawn up of relationships according to types (broader, narrower, related descriptors). There remains however the problem of interpreting the meaning of the links thus established. A compact and logical categorization should be sought whereby the relationship is denominated - e.g. "thing/part, thing/property, process/agent, thing/application, etc.". - and properly defined. Informal relationships may also obtain, usually under the guise of "see" or "see also" references. These in turn are to be listed and tabulated, and if possible, interpreted similarly as formal links.

Step 6 The results of the source thesaurus analysis may then be translated into target formulations.

Summary of first draft establishment

The following operations are required to elaborated a first draft of a multilingual thesaurus:

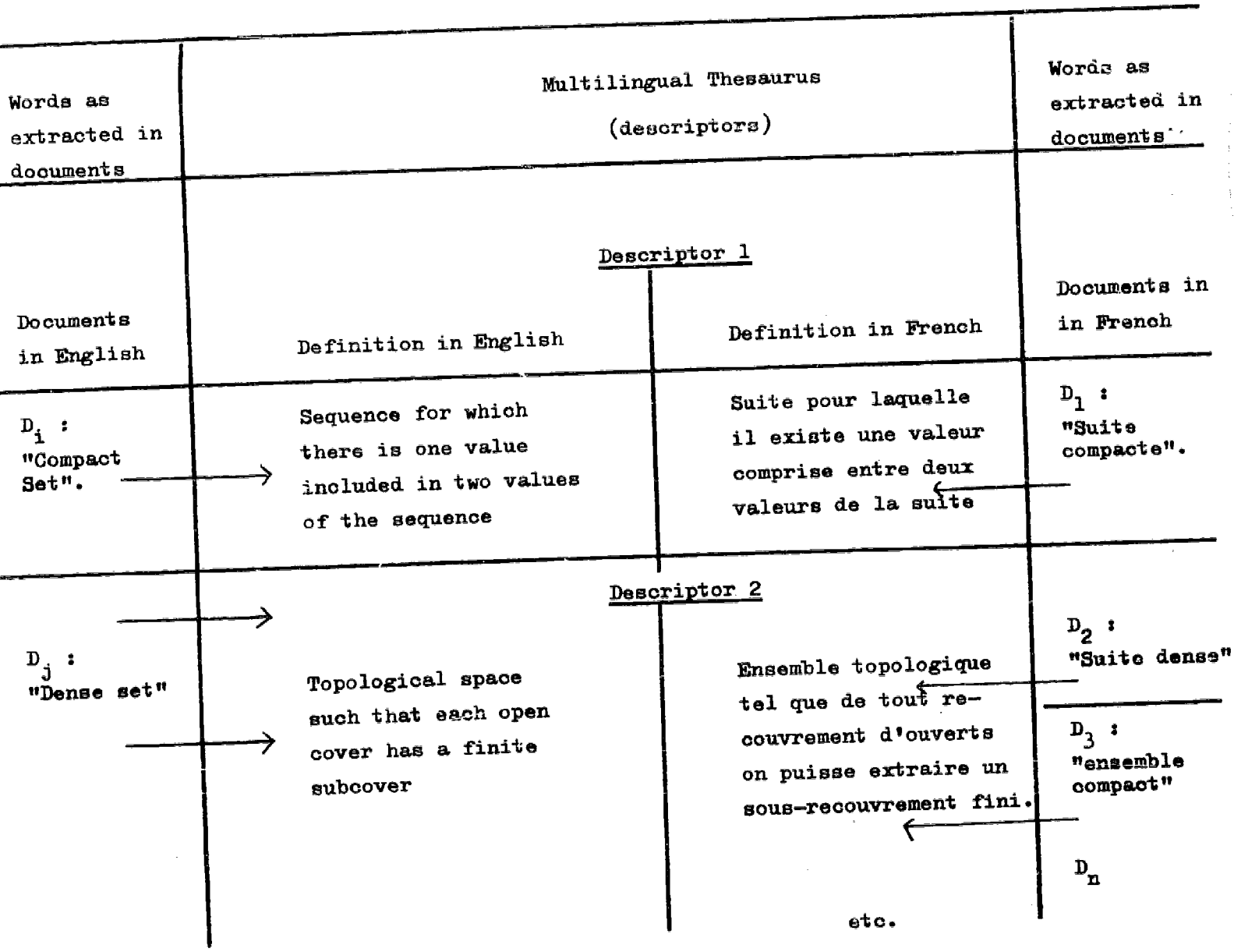

Figure 3: Complex equivalences between words and definitions of descriptors (examples in Mathematics).

→ : relationships between expressions and definitions

Notation

Step 1: Target natural language descriptor denomination symbol notation

Translation

Step 2: Translation of general comments
Translation of definitions and scope-notes

Semantic interpretation

Step 3: Elaboration of new target definitions

Step 4: Listing of multiple equivalents

Structural interpretation

Step 5: Listing and interpretation of formal and internal inter-relationships

Step 6: Translation into target version and editing.

3.2. Presentation of the first draft of the target version

3.2.1. Identical, and corresponding tables can be used to represent the source thesaurus and the target versions. Entries are source descriptors, ordered alphabetically in each table, to which target descriptors are made to correspond.

3.2.2. Graphic display of groups of related source descriptors, and their equivalent target versions help point out multiple equivalences and possible inconsistences of structural interpretations.

/...

3.2.3. Alphabetical lists of inexact-matching descriptors should be kept for ulterior scrutiny and checks.

4. Tests and Checks

4.1. The completion of the previous steps provides a thesaurus with which tests can be carried out: corrections and improvements are then possible. However, the applications of the first draft requires that some material be selectively collected.

Step 7: Samples of documents should be collected, selecting in particular those related to multiple equivalence descriptors. This sample should not be restricted to documents formulated in any one natural language but rather reflect the variety met in an actual - or wanted - corpus of an information system.

A number of potential queries, formulated in target natural language should also be collected.

4.2. The fact of multiple equivalence already mentioned - and reducible to the case where one source descriptor is equivalent to more than one target descriptor - may be due to a variety of causes. Two main types may obtain.

4.2.1. There exists, in the field of application, whether it be discipline - or mission-oriented, a major language in which are expressed the concepts and practices of the domain. Those last may either be absent from the target language or expressible through complex statements, combinations of existing terms, etc. The problem will then be to translate definitions. New terms may be either created, or existing ones used as such in the target language.

/...

Step 8: Definitions of the source descriptors are to be expressed in the target language. Use can be made of mono- or multilingual glossaries when they are available; the validity of the new definitions can also be tested with the help of source thesaurus. Provisional notation may use either the source thesaurus form, or any conventionally recognizable formulation. The new definitions should reflect, as the case may be, the combination of specific target language words which have been revealed by indexing sample documents.

4.2.2. The existence of multiple equivalences may be due, on the other hand to the fact that, in a given domain, each natural language conveys specific meanings: notions may be more or less narrowly related to local conditions - of a cultural, historical, national, or linguistic nature. Then one already defined descriptor, in source language, will find its equivalent in target language only through complex periphrastic expressions. The problem is then to create new definitions, and eventually, new terms.

Step 9: Whenever such descriptors are recognized, they should be embodied in especially formulated queries in target language - if possible extracted from the sample collected queries. Using these to retrieve documents judged relevant in sample collection, a list of words in target language is drawn up: these words are selected which seem to point out best what is conveyed by the source descriptor. They can lead to the eventual creation of new terms.

Step 10: The words thus selected are used to build up a list of candidate target descriptors. They are defined, grouped as need be, and their inter-relationships established: "these can be expressed by several means. If codes are used to indicate these relationships, their meaning should always

/...

be made clear. It is quite evident that glossaries and dictionaries may be of help in fulfilling this task.

Step 11: The candidate descriptors are then matched against those found in the already established first draft, and a tentative allocation is noted. Use is made here of the previously defined cat ories of inter-relationships.

4.3. There remains the task of checking more generally the target version of the thesaurus.

Step 12: All the documents in the sample collection are then indexed with the help of the first draft of the thesaurus. Documents which have required the use of candidate descriptors are indexed twice: once with the help of the source descriptors which correspond to the candidates; the second time with the candidate descriptor. Sample queries are then used to retrieve documents and sets of those judged relevant are grouped.

Step 13: Systematic lists are built up between "clue-words" - occuring in groups of documents retrieved in target natural language - and their corresponding target descriptors. Such selective concordances provide admittedly limited but at least operational "definitions" of descriptors. It enables matching a list of words considered as meaningful "equivalents" (non-descriptors) of a descriptor, as they obtain from indexing and retrieving processes, with the definitions arrived at through translations. Corrections and adjustments of definitions are then carried out until maximum congruence is achieved between the basic thesaurus and the target version.

/...

5. Establishment of a stabilized target version

Before the thesaurus has become a working tool a number of decisions have to be taken first on a final notation, then on a presentation of the necessary components for indexing purposes, finally on the tentative, or alternative solutions arrived at during the previous phases of work.

5.1. Notation: it is unrealistic to expect that the same principles of word-notation - i.e. the symbolization of descriptors with the help of natural language words- bearing on word forms, compound expressions, word order, etc. can hold for a variety of natural languages. As in the case of monolingual thesauri, one should strive at ease of comprehension, coherence, economy, etc. in locally varying degrees, relying on the equivalences already built up to relate the many versions of the thesaurus.

5.2. The indexing tool will present the usual lists and repertories common to most thesauri:

5.2.1. an alphabetical list of descriptors, with their definitions, and indications of their formal interrelationships;

5.2.2. a graphic display of the thesaurus formal structure;

5.2.3. a list of descriptor/natural language words recognized equivalences, etc., depending on local conditions of information processing

5.3. The internal store

Some information will make up the internal store of the multilingual equivalences. This internal store has no function in the indexing process. It will have two parts:

/...

5.3.1. lists of all the equivalences established, at the descriptor or natural language levels, between the source and the target versions: whether hand or machine processible, they are necessary for any systematic, and eventually automatic interchange of indexed documents;

5.3.2. the second part will be made up of all the c idate descriptors, their definitions, and information about their tentative allocation and as to the meaning given to descriptor interrelationships. Such data are to be presented in such a way - whether machine-readable or not - that it can be easily disseminated among the multilingual thesaurus users.

6. Up-dating and maintenance

As the use of multilingual thesauri implies an amount of co-operative effort, if only for the different versions not to diverge in time, it will be found necessary to institute channels of communication between various users, and decision-making procedures. Two alternative solutions can be envisaged depending upon the operational context.

6.1. in large scale information agencies the setting up of a permanent team of specialists may be found justified. Their task would be mainly one of centralizing periodical variations coming from thesauri users; to make adequate decisions and have them known.

6.2. in largely decentralized and federated operations, a small secretariat would be entrusted with the tasks of collecting, and disseminating the tentative decisions of practitioners; it would also ensure that qualified representatives of multilingual thesauri users do convene periodically so as to enforce the necessary homogeneity between different versions.

GLOSSARY

Candidate descriptor: concept recognized and defined for which a descriptor is tentatively established.

Equivalences (between descriptors): concepts which are found to be equivalent both through their definitions in two or more natural languages and through the working of the information system (operational equivalence).

Equivalent descriptors: descriptors which are found to be equivalent in their definitions and their operations

Internal store: all the recorded means by which thesauri are made equivalent and which are not used for indexing or retrieving purposes.

Multilingual thesaurus: equivalent versions, in two or more natural languages, of a thesaurus.

Multiple equivalences (between descriptors): non-overlapping definitions between descriptors belonging to two different versions.

Notation: codes and symbols assigned to descriptors for indexing purposes.

Semantic interpretation: meaning given to the relationships that obtain between descriptors.

Source (or basic) thesaurus: thesaurus used as a basis for establishing multilingual versions.

Source descriptor: descriptor belonging to the basic version of the multilingual thesaurus (source thesaurus).

/...

Target descriptor: descriptor belonging to one of the versions - but not to the source one - of the multilingual thesaurus.

Target thesaurus: thesaurus versions in one natural language equivalent to another version in another language, established with the help of the source thesaurus.

Tentative allocation: temporary allocation of a descriptor to a thesaurus, subject to ulterior checks.

Transformation (of descriptors): set of procedures through which descriptors are made operationnally equivalent.